УДК 004.912:316.42:364
DOI: https://doi.org/ 10.17721/3041-2323.2024.172-184

Юзеф КОРБИЧ[1], д-р техн. наук, проф.
ORCID ID: 0000-0001-8688-3497
e-mail: J.Korbicz@issi.uz.zgora.pl

Олексій ШОЛОХОВ[2], канд. фіз.-мат. наук
ORCID ID: 0000-0002-8676-3724
e-mail: gyroalex@knu.ua

Олексій ЗАРУДНИЙ[3], асп.
ORCID ID: 0009-0008-7462-3899
e-mail: oleksii.zarudnyi@gmail.com

Роман КОВАЛЬ[3], асп.
ORCID ID: 0009-0003-3821-3378
e-mail: roman.koval.science@gmail.com

[1]Університет Зеленої Гури, Польща
[2]Київський національний університет імені Тараса Шевченка, Київ, Україна
[3] Інститут телекомунікацій і глобального інформаційного простору НАН України, Київ, Україна

МЕТОДИКА ЗАСТОСУВАННЯ КЛАСТЕРИЗАЦІЇ ТЕКСТІВ НА ОСНОВІ ЛІНГВІСТИЧНИХ ПРАВИЛ ДЛЯ ДОСЛІДЖЕННЯ ПОТРЕБ НАСЕЛЕННЯ У СОЦІАЛЬНОМУ ЗАХИСТІ ТА СОЦІАЛЬНОМУ ЗАБЕЗПЕЧЕННІ

Питання соціального захисту та соціального забезпечення завжди були одними з найактуальніших для всіх без винятку верств суспільства. А в умовах війни ця сфера набула особливого значення, адже від ефективності державної політики соціального захисту та соціального забезпечення залежить не лише добробут громадян і збалансованість розвитку суспільства, а й забезпечення національної безпеки. За період війни обсяги витрат на соціальний захист і соціальне забезпечення зросли суттєво і зростатимуть надалі, незважаючи на обмеженість бюджетного фінансування. Тому особливої уваги потребує цільова спрямованість коштів на соціальний захист і соціальне забезпечення, контроль за адресністю надання державної допомоги. Оскільки в умовах війни проведення соціологічних досліджень значно

ускладнюється, перспективним напрямом стає дослідження онлайн-середовища. Значна кількість населення використовує різноманітні соціальні мережі, цифрові платформи державних установ та організацій тощо. Тому, маючи інформацію з інтернет-джерел, можна дослідити проблеми, які є значущими для різних соціальних груп, проаналізувати настрої та очікування населення.

Запропоновано методику побудови аналітичної моделі дослідження проблем соціального захисту та соціального забезпечення, що потребують особливої уваги з боку держави, із використанням засобів аналізу текстової інформації з інтернет-джерел і побудови моделей класифікації.

Ключові слова: ***кластеризація текстів, лінгвістичні правила, інтелектуальний аналіз даних, соціальний захист і соціальне забезпечення, інформаційна технологія.***

Вступ

Інформаційно-аналітична діяльність в умовах поглиблення цифровізації суспільства стає все більш важливою складовою системи соціального захисту та соціального забезпечення, що також, як зазначають вітчизняні та закордонні фахівці (Valls et al., 2021; Schüring, & Loewe, 2021; Міністерство цифрової трансформації України, н. д.) потребує її постійної модернізації, впровадження сучасних моделей, методів та інформаційних технологій. Впровадження "Єдиної інформаційної системи соціальної сфери" (Про затвердження Положення…, 2021) стало новим кроком до наскрізної діджиталізації системи пенсійного забезпечення та соціального захисту населення. Метою запровадження Системи є "забезпечення цілісної автоматизації процесів у соціальній сфері шляхом оптимізації та розвитку електронної інформаційної взаємодії суб'єктів Єдиної системи, спрямованої на забезпечення прозорості соціальної сфери, цифровізацію ринку соціальної підтримки та підвищення рівня її доступності для осіб, які її потребують" (Про затвердження Положення…, 2021). Тому питання розроблення методик, моделей, інформаційних технологій аналізу текстової інформації з інтернет-джерел є актуальним і має практичне значення (Гладун, & Рогушина, 2016; Литвин, Пасічник, & Нікольський, 2017; Бодянський, 2016). У ході дослідження розглянуто практичну задачу визначення потреби у соціальному захисті та соціальному забезпечення мешканців різних регіонів України та біженців. Для

аналізу текстової інформації використано інструменти SAS Text Miner (SAS Institute, 2012, 2014; Matignon, 2007).

Результати

Перший етап дослідження виконано на матеріалах інтернет-видань, різних за тематикою та аудиторією, державних і недержавних, з яких відібрано 162 (назви джерел і посилання на них представлено в табл.1).

На основі аналізу текстів, щодо питань соціального захисту та соціального забезпечення, розміщених на вказаних інтернет-ресурсах, отримано шість кластерів.

До першого кластера увійшли тексти, які містять питання, пов'язані з пенсійною реформою. Найхарактернішими для цього кластера виявились слова та словосполучення: "реформа", "страхові виплати", "страховий стаж", "обов'язкові пенсійні накопичення".

До другого кластера ввійшли слова та словосполучення, що описують питання нарахування та виплати пенсій і соціальних допомог. Пенсійним фондом України: "своєчасна виплати пенсій", "добровільні внески на пенсійне страхування", "мінімальна пенсія", "індексація пенсій", "підвищення пенсій", "житлова субсидія", "фінансування поточних виплат", "перерахунок пенсій працюючим пенсіонерам".

Таблиця 1

Перелік інтернет-джерел, інформацію з яких використано для аналізу

Номер	Назва джерела	Адреса ресурсу	Кількість текстів
1	Укрінформ	https://www.ukrinform.ua/rubric-society	50
2	Суспільне. Новини	https://suspilne.media	25
3	Сайт міжнародного наукового видання "Фінансово-кредитна діяльність: проблеми теорії та практики"	https://fkd.net.ua	7
4	Газета "Урядовий кур'єр" – офіційне друковане видання Кабінету Міністрів України	https://ukurier.gov.ua/uk/articles	30

Закінчення табл. 1

Номер	Назва джерела	Адреса ресурсу	Кількість текстів
5	Офіційний сайт Київської обласної ради професійних спілок	http://korps.com.ua	5
6	Офіційний сайт Національного банку України	https://knpf.bank.gov.ua	10
7	Офіційний сайт журналу "Forbes Ukraine"	https://forbes.ua	15
8	Сайт електронного видання "Судово-юридична газета"	https://sud.ua	20

Третій кластер узагальнює проблеми соціального захисту внутрішньо переміщених осіб. Найхарактернішими є такі слова та словосполучення, як "ВПО", "ідентифікація", "звільнені території", "виплати переселенцям", "мешканці окупованого Криму", "Всесвітня продовольча програма ООН" "тимчасово непідконтрольні території".

До четвертого кластера ввійшли слова та словосполучення, що описують проблеми, пов'язані з втратами, внаслідок воєнного конфлікту: "військовослужбовець", "поліцейський", "зона бойових дій", "зниклий безвісти", "втрата годувальника", "члени родини загиблого".

Для п'ятого кластера актуальними є питання соціального захисту та соціального забезпечення біженців, зокрема, "пенсія за кордоном", "праця за межами України", "пропорційне обчислення страхового стажу", "страховий стаж, одержаний в інших країнах".

У шостому кластері узагальнено питання, що стосуються постраждалих унаслідок аварії на Чорнобильській АЕС: "аварія", "ЧАЕС", "чорнобилець".

Статистичні характеристики побудованої моделі класифікації на основі лінгвістичних правил, обчислено окремо для тренінгового та тестового наборів даних: співвідношення – 70 % для тренінгу та 30 % для тестування, тобто 114 та 48 текстів відповідно (табл. 2).

Таблиця 2

Статистичні характеристики моделі класифікації досліджуваних текстів

Статистика	Набір даних	
	Тренінговий	Тестовий
TP (True Positive)	30	11
TN (True Negative)	67	26
FP (False positive)	10	6
FN (False Negative)	7	5
MISC, % (частка неправильно класифікованих значень)	15	23
Gini	0,82	0,71
ROC	0,79	0,67

Зображення ROC-кривої для моделі класифікації текстової інформації на основі лінгвістичних правил подано на рис. 1.

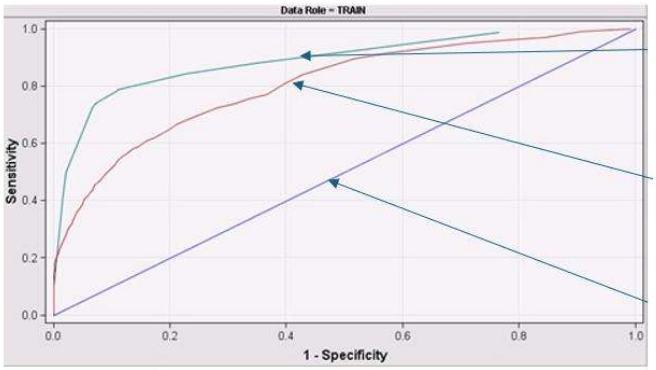

Рис. 1. *ROC-крива для побудованої моделі класифікації на основі лінгвістичних правил*

Побудовані лінгвістичні правила використано для кластеризації текстів новин, що були опубліковані в інтернеті з вересня 2023 р. по вересень 2024 р. Загалом, вивантажено й оброблено близько

10 тис. текстів за тематикою соціального захисту та соціального забезпечення українців.

Після кластеризації текстів, для кожного кластера розраховано кількість текстів, які належать дописувачам із певного регіону. Отримані значення нормовано по шкалі від 0 до 100 за формулою

$$popularity_i = \frac{n_i}{\max(n_i \mid \forall\, i)},$$

де $popularity_i$ – популярність текстів відповідного кластера, для i-го регіону, n_i – кількість текстів по регіону, max $(n_i \mid \forall\, i)$ – максимальна кількість текстів за всіма регіонами. Результати розрахунків представлено у табл. 3.

Таблиця 3

Результати кластерного аналізу текстової інформації з питань соціального захисту та соціального забезпечення по регіонах України

Назва регіону	**Популярність текстів відповідного кластера**					
	Кластер 1. (пенсійна реформа)	Кластер 2 (нарахування та виплата пенсій та соціальних допомог Пенсійним фондом України)	Кластер 3 (проблеми соціального захисту внутрішньо переміщених осіб)	Кластер 4 (проблеми, пов'язані з втратами, внаслідок воєнного конфлікту)	Кластер 5 (питання соціального захисту та соціального забезпечення біженців)	Кластер 6. (питання, що стосуються постраждалих внаслідок аварії на ЧАЕС)

Закінчення табл. 3

Назва регіону	Популярність текстів відповідного кластера					
Вінницька область	94	65	24	72	79	
Волинська область	87	57	20	100	63	
місто Київ	82	49	32	37	26	33
місто Севастополь	-	-	-	-	-	-
Дніпропетровська область	58	39	43	33	14	
Донецька область	27	32	59	37		
Житомирська область	94	73	19	34	62	88
Закарпатська область	67	45	29	40	75	
Запорізька область	58	39	90	30		
Івано-Франківська область	87	66	24	63	72	
Київська область	84	42	28	37	37	100
Кіровоградська область	92	88	32	73	46	
АР Крим	-	1	1	-	-	-
Луганська область		22	8			
Львівська область	73	45	20	60	57	
Миколаївська область	76	70	64	47	18	
Одеська область	32	24	27	13	13	
Полтавська область	75	63	32	73	42	77
Рівненська область	100	64	17	81	100	
Сумська область	92	100	52	43	30	
Тернопільська область	50	56	24		63	
Харківська область	47	35	100	15	9	
Херсонська область	71	62	89			
Хмельницька область	87	55	28	78	73	
Черкаська область	87	47	30	50	55	74
Чернігівська область	81	58	24	50	29	
Чернівецька область	83	31	25	1	61	

Результати аналізу, представлені в таблиці можна візуалізувати використовуючи інструменти SAS Enterprise Guide 7.1 (рис. 2 і 3).

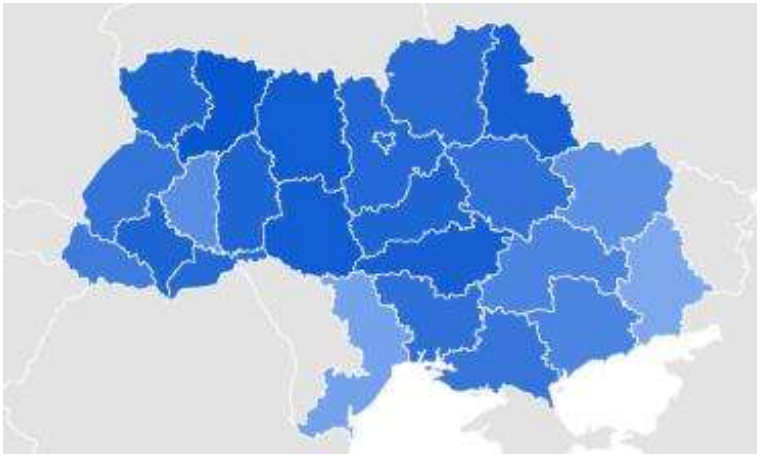

Рис. 2. *Кластер 1 популярності текстів за тематикою "Пенсійна реформа" щодо регіонів України*

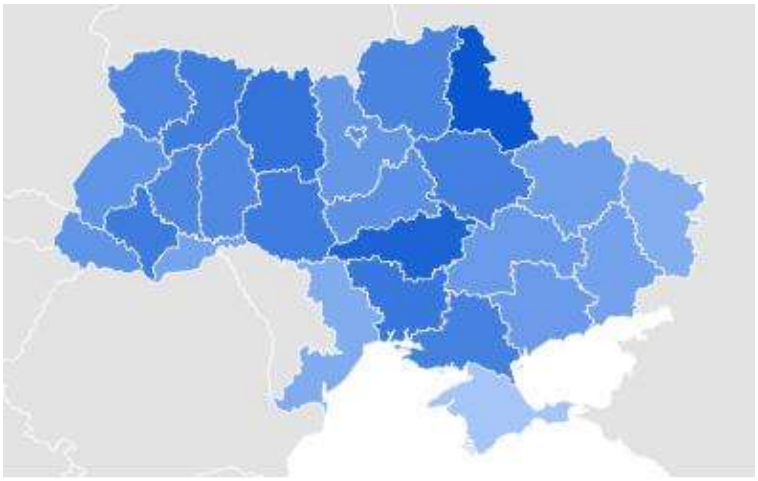

Рис. 3. *Кластер 2 популярності текстів за тематикою "Питання, пов'язані з Пенсійним фондом взагалі" за регіонами України*

Дискусія і висновки

Отримані результати можуть бути в подальшому використані у моделі актуарних розрахунків, під час планування соціальних видатків бюджетів різних рівнів. Пропонований підхід може забезпечувати покращення якості прогнозів у сучасних умовах, коли немає повної інформації про досліджуваний процес чи явище або інформацію спотворено.

Список використаних джерел

Бодянський, Є. В. (2016). *Аналіз та обробка потоків даних засобами обчислювального інтелекту.* Львів: Вид-во Львівської політехніки.

Гладун, А. Я., & Рогушина, Ю. В. (2016). *Data mining: пошук знань в даних.* АДЕФ-Україна.

Грень, Т. Я. (2022). Особливості реалізації політики соціального захисту територій в умовах війни. *Вчені записки ТНУ імені В. І. Вернадського. Серія "Публічне управління та адміністрування"*, 33(72), 6, 81–84. https://doi.org/10.32782/TNU-2663-6468/2022.6/13

Ланде, Д. В. (2014). *Елементи комп'ютерної лінгвістики в правовій інформатиці.* Київ : НДІІП НАПрН України.

Литвин, В. В., Пасічник, В. В., & Нікольський, Ю. В. (2017). *Аналіз даних та знань.* Магнолія 2006.

Міністерство фінансів України. (н. д.). *Видатки на соціальну допомогу.* https://mof.gov.ua/uk/expenditures_on_social_assistance

Міністерство цифрової трансформації України. (н. д.). *Офіційний сайт Міністерства цифрової трансформації України.* https://thedigital.gov.ua

Перебийніс, В. І. (2013). *Статистичні методи для лінгвістів.* Нова Книга.

Про затвердження Положення про Єдину інформаційну систему соціальної сфери. (2021). Постанова Кабінету Міністрів України від 14.04.21. № 404. https://zakon.rada.gov.ua/laws/show/404-2021-п#Text

Смуш-Кулеша, М., Федорова, А., & Мойса, Б. (2022). *Соціальні права в Україні під час війни. Звіт про оцінку потреб.* Рада Європи. https://rm.coe.int/needs-assessment-ua-2/1680a9b408

Шаповалова, Т. (2022). Поняття і зміст соціального захисту та соціального забезпечення населення в сучасній Україні. *Економічний аналіз*, 32(3), 123–130. https://doi.org/10.35774/econa2022.03.123

Шкурко, О. В. (2018). *Види лінгвістичного аналізу тексту.* Університет ім. Альфреда Нобеля.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data.* Springer.

Berry, M. W. (Ed.). (2003). *Survey of text mining I: Clustering, classification, and retrieval.* Springer.

Do Prado, H. A., & Ferneda, E. (Eds.). (2007). *Emerging technologies of text mining: Techniques and applications.* Idea Group Reference.

Find the information that matters using natural language processing (NLP). (n. d.). https://www.sas.com/ru_ua/software/visual-text-analytics.html

Matignon, R. (2007). *Data mining using SAS Enterprise Miner.* https://www.amazon.com/Data-Mining-Using-Enterprise-Miner/dp/0470149019

Schüring, E., & Loewe, M. (Eds.). (2021). *Social protection systems.* Edward Elgar Publishing. https://doi.org/10.4337/9781839109119

Sharma, S., & Jain, A. (n. d.). Role of sentiment analysis in social media security and analytics. *WIREs Data Mining and Knowledge Discovery*, 10(5). https://doi.org/10.1002/widm.1366

SAS Institute. (2012). *Getting started with SAS® Text Miner 12.1.* https://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf

SAS Institute. (2014). *Text analytics using SAS Text Miner: Course notes.* https://documentation.sas.com/?docsetId=tmref&docsetTarget=n1d7r58qug6sefn162cu6cqx0nq4.htm&docsetVersion=14.3&locale=en

Text Cluster Node Results (n. d.). https://documentation.sas.com/?docsetId=tmref-&docsetTarget=n1d7r58qug6sefn162cu6cqx0nq4.htm&docsetVersion=14.3&locale=en

Valls Martínez, M. d. C., Santos-Jaén, J. M., Amin, F.-u., & Martín-Cervantes, P. A. (2021). Pensions, ageing and social security research: Literature review and global trends. *Mathematics*, 9, 3258. https://doi.org/10.3390/math9243258

References

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data.* Springer.

Berry, M. W. (Ed.). (2003). *Survey of text mining I: Clustering, classification, and retrieval.* Springer.

Bodianskyi, Y. V. (2016). *Analysis and processing of data streams using computational intelligence*. Lviv Polytechnic Publishing [in Ukrainian].

Do Prado, H. A., & Ferneda, E. (Eds.). (2007). *Emerging technologies of text mining: Techniques and applications.* Idea Group Reference.

Find the information that matters using natural language processing (NLP). (n. d.). https://www.sas.com/ru_ua/software/visual-text-analytics.html.

Gladun, A. Y., & Rogushina, Y. V. (2016). *Data mining: Search for knowledge in data.*. ADEF-Ukraine [in Ukrainian].

Gren, T. Y. (2022). Features of implementing social protection policy in wartime conditions. *Scientific Notes of TNU named after V. I. Vernadsky. Series: Public Administration and Governance*, 33(72), 6, 81–84 [in Ukrainian]. https://doi.org/10.32782/TNU-2663-6468/2022.6/13

Lande, D. V. (2014). *Elements of computational linguistics in legal informatics.* Kyiv: Research Institute of Legal Informatics [in Ukrainian].

Litvin, V. V., Pasichnyk, V. V., & Nikolskyi, Y. V. (2017). *Data and knowledge analysis: Textbook.* Magnolia 2006 [in Ukrainian].

Matignon, R. (2007). *Data mining using SAS Enterprise Miner.* https://www.amazon.com/Data-Mining-Using-Enterprise-Miner/dp/0470149019.

Ministry of Digital Transformation of Ukraine. (n. d.). *Official website of the Ministry of Digital Transformation of Ukraine* [in Ukrainian]. https://thedigital.gov.ua

Ministry of Finance of Ukraine. (n. d.). *Expenditures on social assistance* [in Ukrainian]. https://mof.gov.ua/uk/expenditures_on_social_assistance

Perebyinis, V. I. (2013). *Statistical methods for linguists.* Nova Knyha [in Ukrainian].

Schüring, E., & Loewe, M. (Eds.). (2021). *Social protection systems.* Edward Elgar Publishing. https://doi.org/10.4337/9781839109119.

Sharma, S., & Jain, A. (n. d.). Role of sentiment analysis in social media security and analytics. *WIREs Data Mining and Knowledge Discovery*, 10(5). https://doi.org/10.1002/widm.1366.

Shapovalova, T. (2022). Concept and content of social protection and social security of the population in modern Ukraine. *Economic Analysis*, 32(3), 123–130 [in Ukrainian]. https://doi.org/10.35774/econa2022.03.123

Shkurko, O. V. (2018). *Types of linguistic text analysis.* Dnipro: Alfred Nobel University [in Ukrainian].

Smush-Kulesha, M., Fedorova, A., & Moisa, B. (2022). *Social rights in Ukraine during the war. Needs assessment report.* Council of Europe [in Ukrainian]. https://rm.coe.int/needs-assessment-ua-2/1680a9b408

SAS Institute. (2012). *Getting started with SAS® Text Miner 12.1.* https://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf.

SAS Institute. (2014). *Text analytics using SAS Text Miner: Course notes.* https://documentation.sas.com/?docsetId=tmref&docsetTarget=n1d7r58qug6sefn162cu6cqx0nq4.htm&docsetVersion=14.3&locale=en

Text Cluster Node Results. (n. d.). https://documentation.sas.com/?docsetId=tmref-&docsetTarget=n1d7r58qug6sefn162cu6cqx0nq4.htm&docsetVersion=14.3&locale=en.

Valls Martínez, M. d. C., Santos-Jaén, J. M., Amin, F.-u., & Martín-Cervantes, P. A. (2021). Pensions, ageing and social security research: Literature review and global trends. *Mathematics*, 9, 3258. https://doi.org/10.3390/math9243258.

Отримано редакцією журналу / Received: 17.09.24
Прорецензовано / Revised: 27.09.24
Схвалено до друку / Accepted: 01.10.24

Józef KORBICZ[1], DSc (Engin.), Prof.
ORCID ID: 0000-0001-8688-3497
e-mail: J. Korbicz@issi.uz.zgora.pl

Oleksii SHOLOKHOV[2], PhD (Phys. & Math.)
ORCID ID: 0000-0002-8676-3724
e-mail: gyroalex@knu.ua

Oleksii ZARUDNYI[3], PhD Student
ORCID ID: 0009-0008-7462-3899
e-mail: oleksii.zarudnyi@gmail.com

Roman KOVAL[3], PhD Student
ORCID ID: 0009-0003-3821-3378
e-mail: roman.koval.science@gmail.com

[1]University of Zielona Góra, Zielona Góra, Poland
[2]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
[3]Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

METHODOLOGY FOR APPLYING TEXT CLUSTERING BASED ON LINGUISTIC RULES TO STUDY THE POPULATION'S NEEDS IN SOCIAL PROTECTION AND SOCIAL SECURITY

Issues of social protection and social security have always been among the most pressing for all segments of society without exception. In times of war, this sphere has acquired special significance, as the effectiveness of state policy in social protection and social security determines not only the well-being of citizens and balanced societal development but also the safeguarding of national security. During the war, expenditures on social protection and social security have increased significantly and are expected to continue growing, despite limited budgetary. Therefore, special attention must be paid to the targeted allocation of funds for social protection and social security, as well as to the control of the proper targeting of state aid

Since conducting sociological research during wartime is significantly complicated, exploring the online environment becomes a promising direction. A large portion of the population uses various social networks, digital platforms of state institutions and organizations, and more. Hence, by analyzing information from internet sources, it is possible to investigate issues relevant to different social groups, as well as assess the sentiments and expectations of the population (Sharma & JainRole, n. d.; Shkurko, 2018; Perebyinis, 2013; Lande, 2014; Find the information that matters using natural language processing (NLP), n. d.; Berry, 2003; Aggarwal, n. d.; Text Cluster Node Results, n. d.; Do Prado & Ferneda, 2007).The work proposes a method of building an analytical model for the study of social protection and

social security problems that require special attention from the state, using means of analyzing textual information from Internet sources and building classification models.

Keywords: ***text clustering, linguistic rules, intelligent data analysis, social protection and social security, information technology.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.